Part II Pathology: Basic Statistics

TJ McKinley

Statistics is concerned with collecting, summarising and manipulating data in order to extract quantitative information about a population of individuals. In practical circumstances it is nearly always impossible to obtain the desired information from the entire population under study, instead we must take a random sample of individuals and produce estimates of the quantities of interest. Good introductory texts are, *Practical Statistics for Medical Research*, by Douglas Altman (1999), and *An Introduction to Statistical Modelling*, by W.J. Krzanowski (1998).

These notes provide statistical background to some of the common exploratory and modelling techniques that can be used to extract information from common types of data; they are intended to be used in conjunction with the corresponding practical worksheets, which show how the methodology can be implemented in Excel.

1 Background - random variables

The fundamental concept underpinning statistics is that measureable quantities are *random*, in the sense that they will vary within a *population*. We denote these quantities *random variables*, and the distribution of all possible values of a random variable in the population can be described by an appropriate *probability distribution*.

Random variables will typically take one of three main forms: either *continuous*, *discrete* or *categorical*. A continuous random variable is one that can take *any* value between certain ranges along the real line. For example heights in a population of students, or the distance that they live away from campus.

A discrete random variable is a measurement that can only take a countable number of values e.g. counts of the number of degree passes, or the number of students taking different modules. A categorical random variable is one that corresponds to particular attributes of an individual, such as gender (male or female), or degree class (fail, third, second or first).

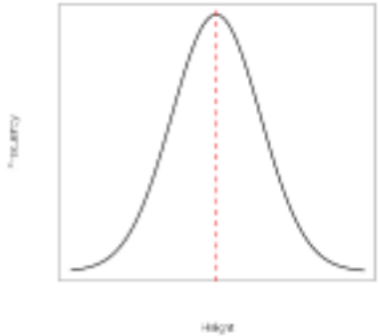


Figure 1: Example normal distribution

Different types of variable follow different types of probability distribution, and it is predominantly the role of statistics to estimate aspects of the underlying distribution. For example, consider a random variable that follows a normal distribution (Figure 1). A normal distribution has a specific form that is determined by two *parameters*, the *mean* and *variance*. A straightforward way of estimating the mean and variance is to use the *sample mean* and *sample variance* (see the next section). Other types of distribution may be governed by different types of parameter (for example see the skewed distributions in Figure 2). In the first instance let's consider some simple descriptive measures to summarise the data.

2 Descriptive statistics

Let's first consider some straightforward methods for summarising and presenting the data in an appropriate manner. This is an important stage, since it can provide useful insights into attributes of the underlying population, and can help to influence the choice of relevant statistical techniques for further analyses.

Let us consider that we are interested in a random variable X, and we have a random sample of n individuals from a population, denoted $X_1, \ldots, X_n$. Three common summary measures are given by the mean, median and mode (Figure 2).

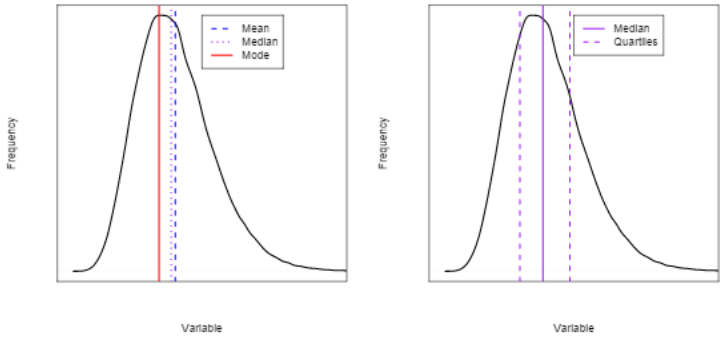


Figure 2: Example summary measures

The *mean* (denoted $\bar{X}$) is given by summing up the values in the random sample and dividing by the number of samples. i.e.

$$\bar{X} = \sum_{i=1}^{n} X_i.$$

The *median* is defined as the middle value, and is given by the $\left(\frac{n+1}{2}\right)^{\text{th}}$ observation, once the data have been sorted into ascending order. Be careful though, since if the $\left(\frac{n+1}{2}\right)^{\text{th}}$ observation lies between two values, then the median must be *interpolated* from the data. The median is more robust to the influence of extreme values in the data than the mean.

A third summary measure is given by the *mode*, simply defined as the most common number. This measure is particularly useful for discrete data.

As an illustrative example let us consider a set of 5 observations from a process, with values 12, 7, 8, 10, 8. Typically we denote *random variables* using capital letters (i.e. X), and observations on those random variables by lower case letters (i.e. x). So in this case we have an observed realisation from the distribution of X, and so the observed mean is given by,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{5} x_i = \frac{(12+7+8+10+8)}{5} = \frac{45}{5} = 9.$$

The median is obtained by first ordering the data, and then finding the $\left(\frac{n+1}{2}\right)^{\text{th}}$ number. So here we are interested in the $\frac{5+1}{2} = \frac{6}{2} = 3^{\text{rd}}$ observation from $(7, 8, 8, 10, 12)$, so the median is 8. The mode is the most common number - so in this case the mode is also 8.

These measures are useful since they provide information about the centrality or location of the population, however they do not tell us anything about the variability. This can impact greatly on the inferences we can draw from the data, so let's consider some common summary measures of dispersion.

Firstly, let us consider the so-called *sample variance* and *standard deviation*. The sample variance is defined as,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The standard deviation is given by the square root of the variance, i.e.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

These measures provide estimates for the extent of dispersion around the *mean* of the distribution. It is worth noting at this point that the sample variance can also be written as

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2,$$

allowing for easier computation by hand. (Note also that we are dividing by $n-1$ here since this is the *sample* variance, if we had information about the entire population we would divide by n.)

For the above data points we have variance

$$s^2 = \frac{1}{4}[(12-9)^2 + (7-9)^2 + (8-9)^2 + (10-9)^2 + (8-9)^2] = \frac{16}{4} = 4,$$

and standard deviation

$$s = \sqrt{s^2} = \sqrt{4} = 2.$$

The variance and standard deviation relate to the spread of the distribution around the mean. An alternative measure of spread is given by the *inter-quartile range* (IQR), which is the difference between the upper and lower *quartiles*. To calculate these measures the data need to be ordered as before, and then split into upper and lower quarters. The lower quartile (Q_1) is given by the $\left(\frac{n+1}{4}\right)^{\text{th}}$ number, and the upper quartile (Q_3) by the $\left(\frac{3(n+1)}{4}\right)^{\text{th}}$ number. (The median can be denoted as Q_2.) The IQR is then $Q_3 - Q_1$. In addition we can also split the data into smaller partitions such as *percentiles* (hundredths) or deciles (tenths) if required. (The lower quartile is the 25^{th} percentile, the median the 50^{th} percentile and the upper quartile the 75^{th} percentile.)

The upper and lower quartiles of the example data are given by the $\left(\frac{n+1}{4}\right)^{\text{th}}$ and $\left(\frac{3(n+1)}{4}\right)^{\text{th}}$, or 1.5^{th} and 4.5^{th} numbers from the ordered list $(7, 8, 8, 10, 12)$. Here we will have to interpolate to obtain the quartiles: Q_1 number lies halfway between the first and second numbers (i.e. 7.5), and Q_3 lies halfway between the fourth and fifth numbers (i.e. 11).

Two other useful measures are given by the dimensionless quantities *skewness* and *kurtosis*. The former takes a positive value if the distribution has an asymmetric tail extending toward more positive values, and a negative value if the distribution has an asymmetric tail extending toward more negative values. The further from zero the value of the skewness, then the larger the extent of asymmetry.

Another way of assessing skewness is to consider the remark made earlier about the mean being more susceptible to 'extreme' data points than the median. A good indication of skewness can be therefore be obtained by considering that negatively skewed distributions have mean<median, symmetric distributions have mean=median, and positively skewed distributions have mean>median (though this is not very robust and needs to be used with care).

Kurtosis gives a measure of the relative 'peakedness' or flatness of a distribution in comparison to the normal (Gaussian) distribution. A value greater than zero indicates a more peaked distribution, and a value less than zero to a more flat distribution. We will not give analytical details here.

An simple way of visualising the shape of the underlying distribution is to use a box-and-whisker plot (shown in Figure 3). The lower bound of the box is the lower quartile, the middle line is the median and the upper bound of the box represents the upper quartile. The whiskers extend the range of the data. (These are usually the maximum and minimum points, though it varies from package to package. In this case they show the most extreme value that is less than $1.5 \times$ the IQR from the median. It then highlights points that are beyond this range.) You can see from the plot that there seems to be some positive skew in the data.

For large data sets it is often useful to group the data into a *frequency distribution*, and derive summary measures from the grouped data. Consider that we have a random sample of individuals from a particular distribution, and that we can group these into a series of classes. An example of a measurement for a random sample of 100 individuals is given in Table 1.

Since we are losing information about individual values by grouping them together, in order to estimate the summary statistics for the grouped data we have to assume that the individuals within each class are evenly spread throughout the interval. We can therefore treat the *centre* of each interval as a good approximation for all values within the group

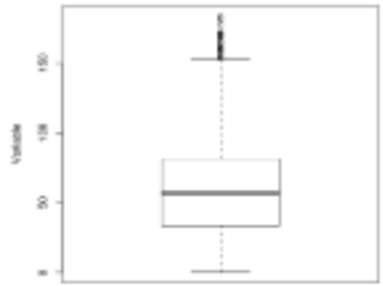


Figure 3: Example box-and-whisker plot

Value (x)	Frequency
$596.5 \leqslant x < 597.5$	1
$597.5 \leqslant x < 598.5$	16
$598.5 \leqslant x < 599.5$	24
$599.5 \leqslant x < 600.5$	8
$600.5 \leqslant x < 601.5$	41
$601.5 \leqslant x < 602.5$	10

Table 1: Example interval data

when calculating the required summary measures. For the above example we have,

$$\begin{aligned}\bar{x} &= \frac{1}{100} \times [(1 \times 597) + (16 \times 598) + (24 \times 599) + \\ &\qquad (8 \times 600) + (41 \times 601) + (10 \times 602)] \\ &= 600.02,\end{aligned}$$

and,

$$\begin{aligned}s^2 &= \frac{1}{99} \times \left\{[(1 \times 597^2) + (16 \times 598^2) + (24 \times 599^2) + \right. \\ &\qquad \left. (8 \times 600^2) + (41 \times 601^2) + (10 \times 602^2)] - 100 \times (600.02^2)\right\} \\ &= 1.798.\end{aligned}$$

The median is given by the 50.5^{th} observation, which again must be interpolated. Here the lower bound for the interval containing the 50^{th} and 51^{st} observations is 600.5, and the 50.5^{th} observation is the 1.5^{th} observation (out of 41) in the interval. Since we are assuming that the observations within each interval are evenly spread, then this observation lies $\frac{1.5}{41}^{\text{th}}$ along the interval $600.5 \leqslant x < 601.5$. So the median is $600.5 + \frac{1.5}{41} \times 1 = 600.54$ (since the class width is equal to one).

Grouped data can be represented by a histogram (Figure 4) in which the area under each bar is equal to the frequency of the value in the interval (i.e. the height of each bar is the frequency divided by the class width, width is equal to the class width and the bar is centred on the mid-point of the group). This is another way of providing information about the shape of the underlying distribution.

3 Interval estimation

We can produce point estimates of quantities such as the mean, median and variance, but often it is of interest to ask: how confident are we in the accuracy of our point estimates? To assess this we can generate *confidence intervals* for our estimates. A (1- α)% confidence interval represents an interval for which we have a (1-α)% probability that the interval contains the *true* mean. (α is known as the *significance* level - typically set to 0.05, or 5%.)

We can produce CIs for any estimated values, but as an illustration consider that we wish to produce a confidence interval for the *mean* of a normally distributed random variable. If we assume the population mean and variance is *unknown*, then we have to estimate them using the *sample* mean ($\bar{x}$) and variance (s^2) from a sample of n individuals. A $(1-\alpha)$% confidence interval for the mean is given by:

$$\bar{x} \pm t_{1-\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}},$$

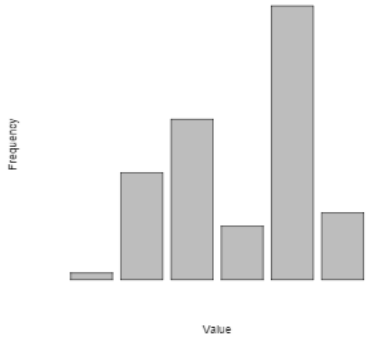


Figure 4: Example histogram for Table 1

where $t_{1-\frac{\alpha}{2},n-1}$ is the $(1-\alpha/2)^{\text{th}}$ percentile of a t-distribution on $n-1$ degrees-of-freedom (Figure 5). This quantity is usually calculated automatically or can be generated in most standard statistical packages. $\frac{s}{\sqrt{n}}$ is known as the *standard error* of the mean. An important point to note is that as the sample size n increases the standard error of the mean decreases, so for very large samples the size of the confidence interval will decrease. Conversely, if the standard deviation of the population gets large, then the standard error will increase and hence so will the size of the confidence interval.

This is important because the same underlying theory applies when we develop hypothesis testing procedures for assessing differences between aspects of two or more populations. This will be discussed in more detail in the next section. As an example consider the data we used before: 12, 7, 8, 10, 8. Assuming this is normally distributed (which it may not be, though it's difficult to tell with very small samples like this) then we know that $\bar{x} = 9$ and $s = 2$. The 97.5^{th} percentile of a t-distribution on 4 degrees-of-freedom is 2.78, so our 95% confidence interval for the mean is:

$$\begin{aligned}&9 \pm 2.78 \times \frac{2}{\sqrt{5}} \\ =\ &(6.51, 11.49)\end{aligned}$$

Due to the fact that we have a small sample size the confidence interval is quite wide, even though the variance is reasonably small in comparison to the mean. Other forms of confidence interval exist for other measures. We can generate CIs for the variance, median, proportions and so on, though the form of these changes depending on the *sampling*

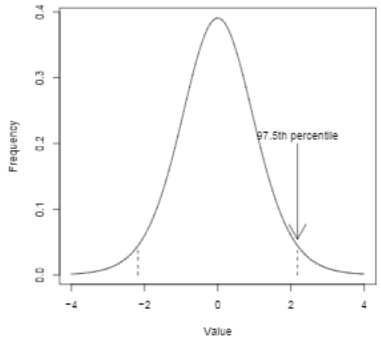


Figure 5: 97.5^{th} percentile of a t-distribution

distribution of the estimate. For the sample mean, $\bar{x}$, the sampling distribution is t-distributed, but for the sample variance, s^2, it follows a chi-squared distribution. Details of other forms of CI will not be given here, suffice to say that you need to be careful to get the correct form (though most statistical packages will automatically calculate these for you).

4 Hypothesis testing

Sometimes you may wish to compare measurements between two or more distinct groups. For example you may wish to compare the means between two populations, to assess whether they are 'different' in some sense, and to assess what that difference is (the *effect* size). Confidence intervals can often be used to that effect, since they can be used to represent the effect size plus the degree of accuracy that we have in the estimated value. However we can also use the notion of *hypothesis testing* to test for *statistically* significant differences between groups.

A hypothesis test is conducted by formulating a *null* hypothesis (H_0) and an alternative hypothesis (H_1). We can then generate a test statistic, that follows a particular sampling distribution under the null hypothesis. We then calculate the value of the test statistic for the observed data, and compare that to the expected sampling distribution. If the observed test statistic lies in the *extremes* of this distribution we say that we *reject* the null hypothesis at the (1-α)% significance level.

As an example here we will assume that you want to test for a difference between the *means* for two populations. To formulate a hypothesis test for comparing two means we have:

$$H_0 : \mu_1 = \mu_2 \qquad \text{vs.} \qquad H_1 : \mu_1 \neq \mu_2. \tag{1}$$

That is we are comparing the null hypothesis that there is no difference between the population means against the alternative hypothesis that the means are different (this is called a two-sided test; one sided tests i.e. $H_1 : \mu_1 < \mu_2$ can be evaluated but must be used with caution). Under the null hypothesis in (1) we can build a test statistic T that follows a t-distribution on $n-1$ degrees-of-freedom (seem familiar?). If our observed statistic $T > t_{1-\frac{\alpha}{2},n-1}$ or $T < t_{\alpha/2,n-1}$ (e.g. it lies at either end of the distribution in Figure 5) then we *reject* the null hypothesis, otherwise we say that the test statistic is consistent with the null hypothesis. In this case we say that the observed p-value is less than α.

Different test statistics can be generated for different hypothesis tests. However there are some important points to note about hypothesis testing:

- P-values measure *precision*, not *effect size*. Just because a difference is *statistically* significant does not mean it is *biologically* significant. It is for this reason that reporting confidence intervals is useful (or at least reporting the observed effect size with a p-value, with a comment about the biological significance of the effect size).
- Hypothesis tests work by building evidence *against* the null hypothesis, if p$< \alpha$ we can reject the null hypothesis, but if p $\geq \alpha$ we *cannot* say that H_0 is true. The test statistic T for comparing means is given by:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

  where n_1 and n_2 are the sample sizes for groups 1 and 2, and s is the *pooled* sample standard deviation. In the same way as confidence intervals, as the sample size increases the denominator decreases, and hence T will get larger and larger. This allows you to be able to detect smaller and smaller effect sizes. Given a large enough sample *any* (biologically negligible) effect size can be deemed statistically significant. This is another reason why reporting the effect size is paramount.)

As stated, different tests exist that depend on the question and the form of the data. Some common examples are below:

- Comparing **means** between multiple (or two) groups: **ANOVA** (**t-test**).

- Comparing **locations**[1] between multiple (or two) groups: **Kruskal-Wallis** (**Mann-Whitney**).
- Looking for a **linear** association between two continuous/discrete variables: **correlation** coefficient.
- Looking for an association between two categorical variables: **chi-squared test**.

5 Bivariate statistics

In practice we are often interested in measuring two or more continuous variables and assessing whether they are related in any way. For example, we might be interested in whether there is a relationship between age and cholesterol in a health study, or between the annual yield of a particular crop and the annual rainfall.

Usually one variable is viewed as the *response* variable (Y say), and the other as an *explanatory* variable (X say). In this case the question arises: is there a relationship between Y and X and to what extent can we predict Y given knowledge of X?

Given a random sample (X_i, Y_i) for $i = 1, \ldots, n$ individuals, some useful bivariate descriptive statistics are:

$$\bar{X} = \sum_{i=1}^{n} X_i,$$

$$\bar{Y} = \sum_{i=1}^{n} Y_i,$$

$$S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2, \qquad S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

and

$$S_{XY} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

Hence the variances of X and Y are given by

$$Var(X) = \frac{1}{n-1} S_{XX} \quad \text{and} \quad Var(Y) = \frac{1}{n-1} S_{YY},$$

respectively. The *covariance* of X and Y is given by,

$$Cov(X, Y) = \frac{1}{n-1} S_{XY}.$$

[1] akin to comparing medians

Age (months)	Height (cm)
18	76.1
19	77
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5

Table 2: Table of values for height and age of a sample of children

[Note that S_{XX} can also be written as $(\sum_{i=1}^{n} X_i^2) - n\bar{X}^2$, S_{YY} as $(\sum_{i=1}^{n} Y_i^2) - n\bar{Y}^2$, and S_{XY} as $(\sum_{i=1}^{n} X_i Y_i) - n\bar{X}\bar{Y}$.]

The covariance is a measure of linear relationship between two variables, and can be used to calculate Pearson's correlation coefficient,

$$corr(X, Y) = \rho = \frac{Cov(X, Y)}{\sqrt{Var(X) Var(Y)}},$$

which assesses the strength of this relationship. This takes values such that $-1 \leqslant \rho \leqslant 1$, where $\rho = 0$ corresponds to no linear relationship, and values of 1 and -1 correspond to a perfect linear relationship. The sign of ρ indicates which direction the association lies. Note the importance of the word *linear* here; having $\rho \approx 0$ does not mean that there is no relationship between two variables, simply that they are not *linearly* related. Also there may be times when there are spurious correlations, possibly due to outlying points in the data for example. It is important to plot the data to assess potential anomalies in addition to just providing summary measures.

Take a look at the set of data shown below (Table 2) for the heights of various children[2] compared to their age:

Producing a plot of the data (Figure 6) shows that there may be a linear relationship between these two variables. We can assess this by producing Pearson's correlation coefficient for age (X) and height (Y).

[2] in the actual data set these numbers represent the *mean* height in each age group. Here we will consider them to be individual children

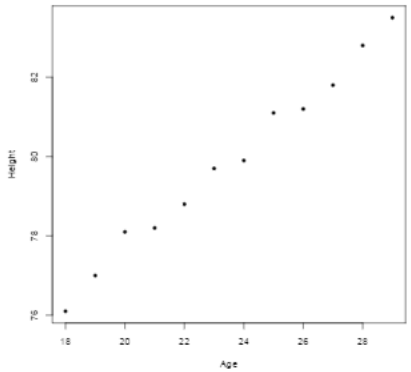


Figure 6: Plot of height against age

To do this we need to know S_{XX}, S_{YY} and S_{XY}. So,

$$\bar{x} = \frac{1}{12} \times (18 + 19 + 20 + 21 + 22 + 23 + 24 + 25 + 26 + 27 + 28 + 29) = 23.5,$$

and

$$\bar{y} = \frac{1}{12} \times (76.1 + 77 + 78.1 + 78.2 + 78.8 + 79.7 + 79.9 + 81.1 + 81.2 + 81.8 + 82.8 + 83.5) = 79.85.$$

This gives,

$$S_{XX} = (18^2 + 19^2 + 20^2 + 21^2 + 22^2 + 23^2 + 24^2 + 25^2 + 26^2 + 27^2 + 28^2 + 29^2) - 12 \times 23.5^2 = 143,$$

$$S_{YY} = (76.1^2 + 77^2 + 78.1^2 + 78.2^2 + 78.8^2 + 79.7^2 + 79.9^2 + 81.1^2 + 81.2^2 + 81.8^2 + 82.8^2 + 83.5^2) - 12 \times 79.85 = 58.31,$$

and

$$S_{XY} = [(18 \times 76.1) + (19 \times 77) + \ldots + (29 \times 83.5)] - 12 \times 23.5 \times 79.85 = 90.8.$$

Pearson's correlation coefficient is then given by

$$\rho = \frac{90.8}{\sqrt{143 \times 58.31}} = 0.9944 (4dp).$$

Indicating a strong positive linear relationship between height and age.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.994366098
R Square	0.988763937
Adjusted R Square	0.987640331
Standard Error	0.25596383
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	57.65482517	57.65482517	879.9914612	4.42807E-11
Residual	10	0.655174825	0.065517483		
Total	11	58.31			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	64.92832168	0.508410245	127.7085235	2.12668E-17	63.79551306	66.06113029	63.79551306	66.06113029
X Variable 1	0.634965035	0.021404771	29.66466003	4.42807E-11	0.587272234	0.682657836	0.587272234	0.682657836

5.1 Simple linear regression

If we believe that two variables, X and Y may be linearly related, then we can fit a simple linear regression model to assess the nature of the relationship. In this way we can predict the value of Y for new individuals based on their observed value of X.

In the case where there is a perfect linear relationship, then for an individual i we have the response variable $Y_i = \beta_0 + \beta_1 X_i$. In reality relationships are rarely perfectly linear, and we will observe some variation. In this case we have:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

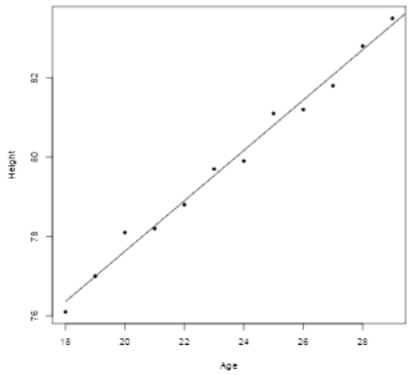
The departure term ϵ_i represents the deviation of individual i from the mean regression line (given by $\beta_0 + \beta_1 X_i$). Here ϵ is drawn from a probability distribution, which in simple linear regression we assume to be normally distributed with a zero mean i.e. $\epsilon \sim N(0, \sigma^2)$ (where σ^2 is the variance).

In regression analysis we are interested in estimating the values of β_0 and β_1 from the data e.g we wish to fit a line of best fit through the data points, with *intercept* β_0 and *slope* β_1. We make the assumptions that X is measured precisely, that the Y_i are independent from each other, and that the expected value (mean) of Y at a point $X = x$ is $E(Y) = \beta_0 + \beta_1 x$. We also assume that the variance is constant along the regression line, so $Var(Y) = \sigma^2$.

The method generally used for estimating the regression parameters β_0 and β_1 is known as *least squares*. For the purposes of this tutorial we do not provide details here. Instead these techniques are implemented in Excel through the 'Analysis ToolPak'. More important here is being able to interpret the output. For the age vs. height data in Table 2 Excel provides us with a summary table of output values:

This has given us an estimated intercept (β_0) parameter of 64.93 and a slope parameter (β_1) parameter of 0.63. If we plot the 'fitted line', $y = 64.93 + 0.63x$ against our original

values we get the output in Figure 5.1.



However it is possible to fit a straight line through any set of points, and of importance in any regression analysis is assessing how well the model 'fits' the data. The *significance* of the regression is quantified by the value of the F-statistic, which tests the significance of the regression against the null model of no association. In the above example this is given by 879.99. This can be compared against an F-distribution and a p-value obtained such that the regression is deemed significant if p is less than a pre-determined value (usually 0.05 for 5% level of significance). Here $p = 4.42 \times 10^{-11}$, so the regression is highly significant.

Another key measure provided by the regression output is the 'R Square' value. This is known as the *coefficient of determination*, and is equal to the square of the correlation (hence $0 \leqslant R^2 < 1$). It is a measure of how well the explanatory variable explains the variation in the response. As a simple rule-of-thumb $R^2 > 0.8$ indicates a 'good' fit to the data, and a value of $R_0 \approx 0.8$ can be interpreted as, 'the variable X explains approximately 80% of the variation in Y'.

Another useful plot is to produce a scatterplot of the residuals (the error around the regression line). This allows us to see whether

- whether the systematic part of the model is correct (i.e. is the relationship truly linear?).
- Whether the variance is constant along the regression line.

We can also produce a normal probability plot, which helps to assess the normality of the data. If the data are truly normal then we expect this plot to be a straight-line. Examples of these are given in the practical.

So if we have a good fit to the data we can then look at the individual parameter estimates and make inferences about the relationship between the response and explanatory variables. Here we have an intercept of 64.93 and a slope parameter of 0.63. This suggests that the mean height of children when they are born is 64.93cm, and that this increases linearly at a rate of 0.63cm per month thereafter.

These are *mean estimates* of the regression parameters however (a change in data will produce a different estimate). An additional measure associated with these values is their standard error, corresponding to a measure of how 'close' the parameter estimates are to their 'true' values. We can test the significance of a parameter estimate to see whether it is statistically significantly different to zero by using a t-test, producing a p-value. If this p-value is less than a pre-determined value (typically $p < 0.05$), then we can say that the parameter is statistically significantly different from zero at the 5% level.

A particularly useful attribute of regression models, other than being able to explain the nature of relationships between variables is that they can be used to predict the response for unobserved individuals. As an example consider that a new child enters the study, aged 25.5 months and we want to know what the 'best' prediction is for their height given their age. Taking $x = 25.5$ and plugging it into the regression equation $y = 64.93 + 0.63x$ gives us a predicted height of 80.995cm. Care must be taken when predicting to values outside of the data range however, since knowledge of how the data behaves outside of this range is unknown (i.e. it may no longer be linear). Predicting outside of the range is known as *extrapolating* the data.